INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY. CABLE CENTRATOM TRIESTE

SMR.764 - 4

RESEARCH WORKSHOP ON CONDENSED MATTER PHYSICS
13 JUNE - 19 AUGUST 1994

MINIWORKSHOP ON
"NONLINEAR TIME SERIES ANALYSIS"
8 - 12 AUGUST 1994

"Neural Networks and Time Series "

Anderas WEIGEND
Computer Science Department
University of Colorado, P.O. Box 430
Boulder, CO 80309-0430
U.S.A.

These are preliminary lecture notes intended only for distribution to participants

CONNECTIONIST MODELING OF TIME SERIES

ANDREAS WEIGEND

DEPARTMENT OF COMPUTER SCIENCE
AND INSTITUTE OF COGNITIVE SCIENCE

UNIVERSITY OF COLORADO AT BOULDER

ITP TRIESTE, AUGUST 1994

STATE SPACE EMBEDDING

- **Express future value as function of past values (Yule, 1927)**

  ... and then just play *"connect the dots"*

- **Assumes system is time invariant (stationary)**

  "Forget the mystery, buy the history"

- **Questions (whatever model is used):**
  - **What time scale of observations?**
    (sampling time)
  - **How far to predict? Direct? Iterated?**
  - **How many past values are necessary/best?**
  - **How to preprocess series?**
    (log, differentiate...)
  - **What (other) inputs are useful?**
    (subset selection)
  - **What cost function should be minimized?**
    (error model)
  - **What other outputs are useful?**

SPACE OF TIME SERIES IS LARGE

- Symbol sequences
- Continuous variables.
  - dynamical systems (nonlinear DEQ)
  - maps
  - (almost) random walks

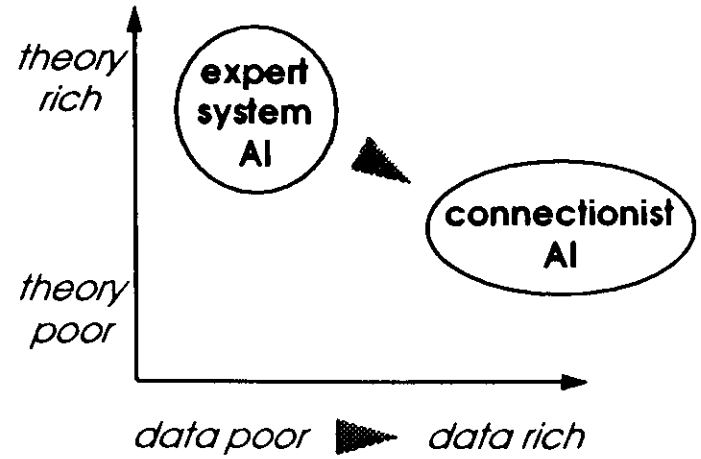
Start from data

vs.

start from assumptions

TRADEOFFS IN MODELING

- bias – variance
- deterministic – stochastic
- strong/narrow – weak/broad



- noise – nonstationarity

PARADIGM CHANGES

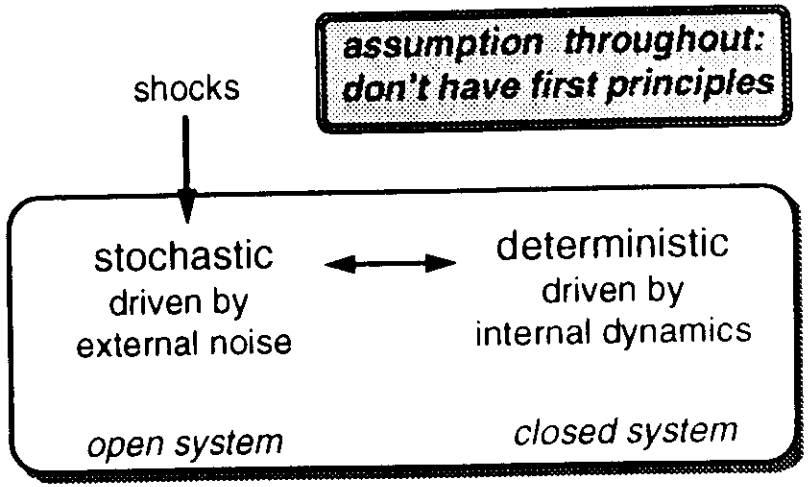
Dimensions of the "Time Series Space"

- **stochastic --> deterministic**
- **linear --> nonlinear**
- **prediction --> characterization**

Strong models --> Weak models

- **data poor & theory rich --> data rich & theory poor**
  - **the more flexible the model, the harder the evaluation**
- **in-sample error --> out-of-sample error**
  - **models so flexible that they overfit, i.e., do very well on the training data but do not generalize well to novel test data**
- **need to always set evaluation data apart not used in the fit**

STOCHASTIC vs. DETERMINISTIC



LINEAR ↔ NONLINEAR

Global Linear Models

- Interpretation (relatively) easy
  - as linear filter
    - frequency response, autocorrelation
    - response independent of amplitude (superposition)
    - no "coupling" between the various inputs
  - as surface fitting
    - one hyperplane in state space
- Model selection (relatively) easy
- Needs to be excited by noise
  - often good in very noisy systems (e.g., economics)
  - but what if systems is low-noise and nonlinear?

Local Linear Models and Nonlinear Models

LOCAL LINEAR MODELS

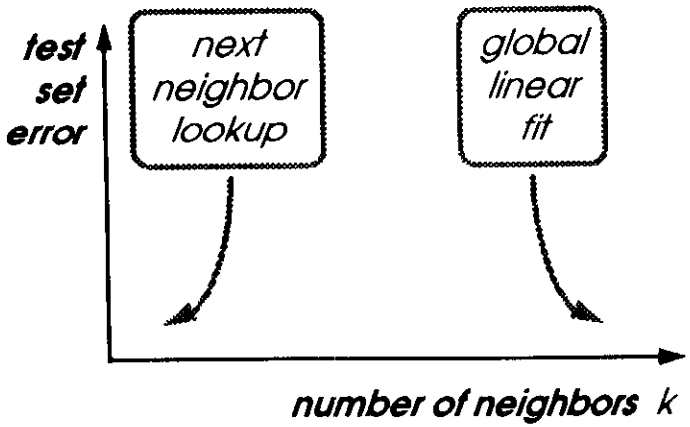
Global linear models

Local linear models

- Local in state space (not in time!)
- E.g., vary size of neighborhood --> characterization
  - DVS Plots = "Deterministic vs. Stochastic", (Martin Casdagli)
  - Bias-variance trade-off
- Extreme case: look-up closest value
  - Instance-based learning
- Want more points in state space? Interpolate in time space !
  - Fill in points of manifold, Tim Sauer

BIAS - VARIANCE TRADEOFF

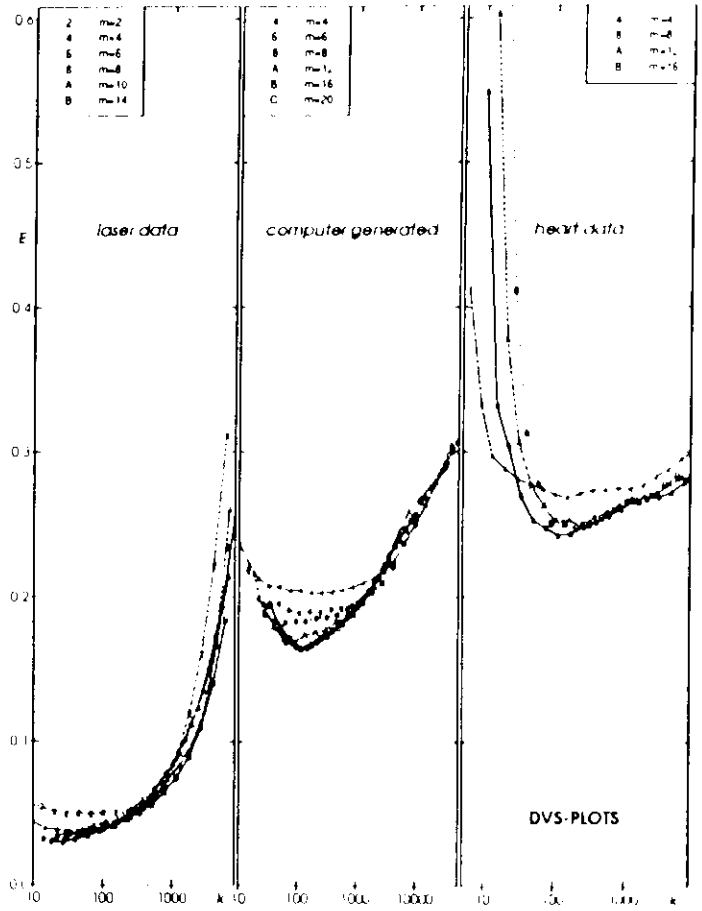
- construct family of local linear models
- vary size of neighborhood
- vary number of lags
- plot test-error (out-of-sample error) as function of number of neighbors



DVS-plots: deterministic vs stochastic

DVS PLOTS

from Casdagli and Weigend (1994)



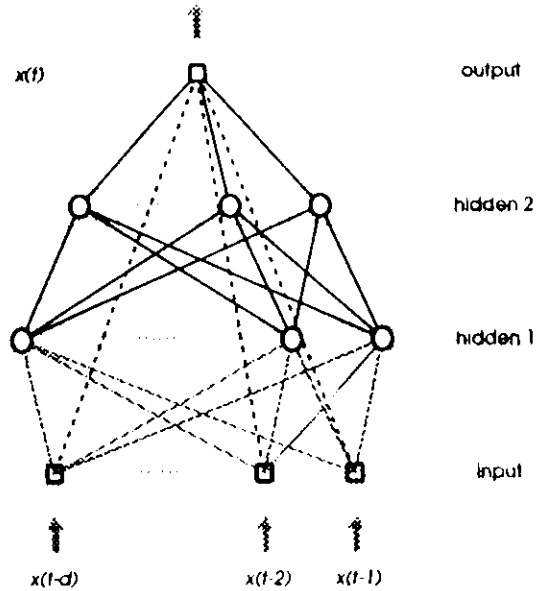
NONLINEAR MODELS

- Large class of models; focus here one feed-forward networks
- Essentially nonlinear regression (after state space embedding)
- Very flexible: neural nets often more parameters than data points
- Model comparison can become very hard
- Summary:
  - Understanding (explicit model) great when it works...
  - ...but learning (implicit model / emulation) broader
  - Easy to make predictions...
  - ...but how good are they?

Understand the error sources and predict the uncertainty of the prediction.

HOW TO PUT TIME/MEMORY INTO A NETWORK

- memory only at input
  - explicitly give past values at input



- distributed memory
  - replace all weights by tapped delay lines
- recurrent networks

PREDICTING IMPREDICTABILITY

- Motivation
  - risk
  - combine predictors (Markowitz)
  - find regions of low uncertainly
    "profit boxes"

1. Local error bars (confidence intervals)

2. Local error bars due to chaos only

3. Probability density of next value

4. Probability density (uncertainly from sampling errors only)

5. Monte Carlo (inject noise)

Local means: 

1. PREDICTING LOCAL ERROR BARS (CONFIDENCE INTERVALS)

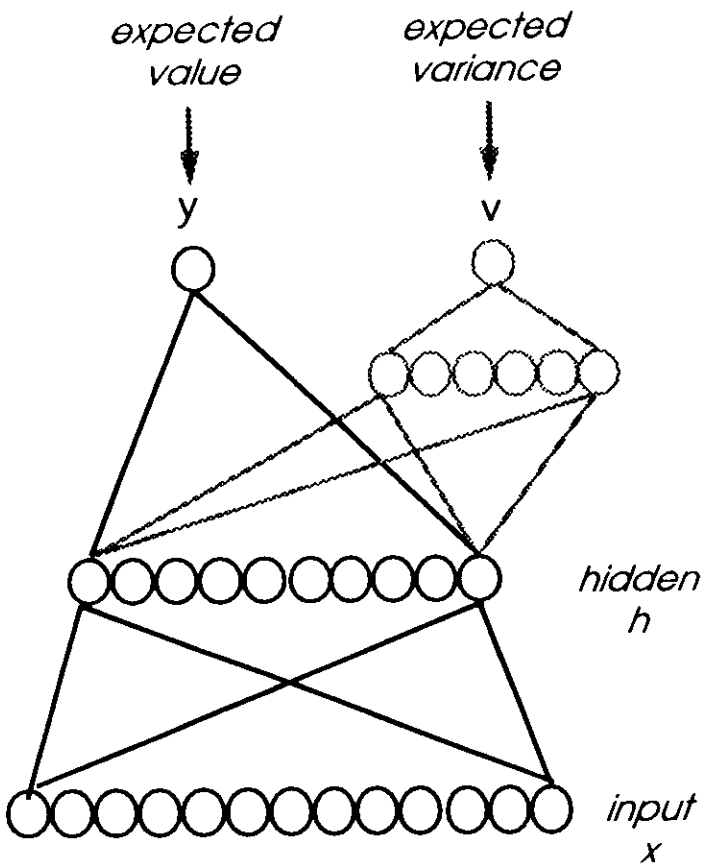
with Dave Nix

ftp.cs.colorado.edu:/Time-Series/error-bars.ps

- Theory / Tricks
  - Maximum Likelihood framework
  - Interpretation as weighted regression
  - Three phases in training
- Examples
  - computer generated data (regression where amount of noise added depends on location)
  - Santa Fe laser data
- Outlook
  - Applicable in regression with any error model with two parameters

    e.g., hyperbolic distribution: it parametrizes Gauss (Euclidean), Laplace (absolute), and exponential errors
  - Also for classification

ARCHITECTURE FOR LOCAL ERROR BARS



COST FUNCTION AND UPDATE RULES FOR LOCAL ERROR BARS

- **Predict error bars**
  - **assume each observed data point d was actually generated from a Gaussian whose mean y and width σ depend on the input**
  - **maximum likelihood (with early stopping)**
  - **estimate y and σ =sqrt(v)**

error model: Gaussian(data|model)

$1/\sqrt{2\pi v(x_l)}\ \exp[-(d_l - y(x_l))^2/2v(x_l)]$

cost function: -log(Gaussian(data|model))

$C = 0.5 \sum_l (d_l - y(x_l))^2/v(x_l) + \ln v(x_l)$

weight change to y unit

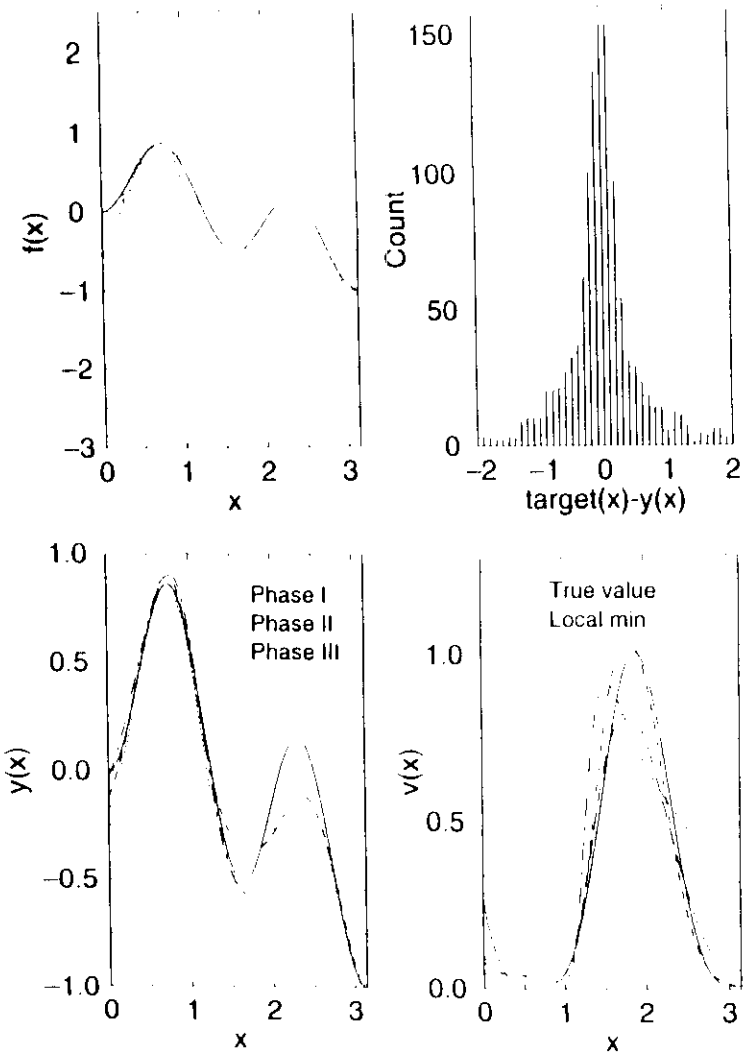
$\Delta w_{yj} = \eta\ 1/v(x_l)\ [d_l - y(x_l)]\ h_j(x_l)$

weight change to v unit

$\Delta w_{vk} = \eta\ 1/2v(x_l)\ [(d_l - y(x_l))^2 - v(x_l)]\ h_k(x_l)$

- **interpretation as weighted regression**

COMPUTER GENENERATED DATA

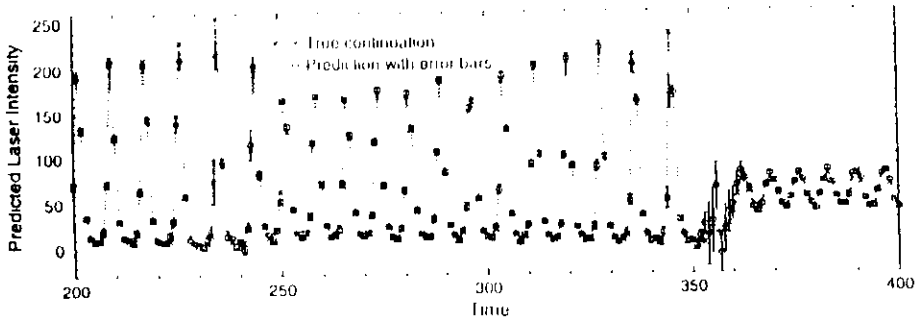


SANTA FE DATA SET A: LASER

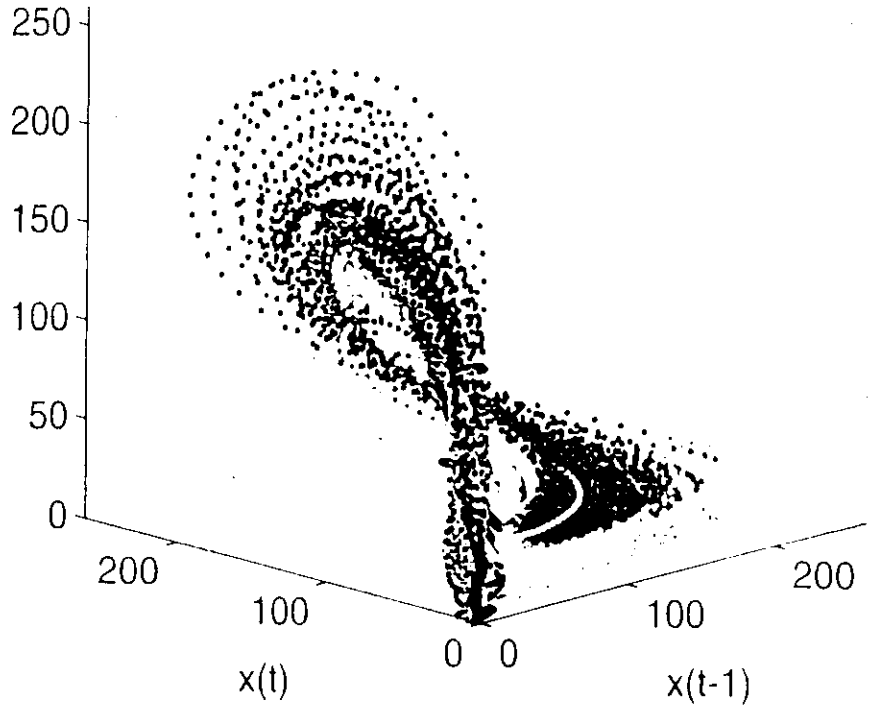
- Description
  - "Univariate time record of a single observed quantity measured in a physics laboratory experiment."
- Origin
  - intensity of far-infrared laser
  - stationary (sampling time 80 nsec)
  - clean (signal to noise ratio 300:1, 8 bit)
  - deterministic chaos
  - dimension around 2.1
- Provided 1000 points

  continuation kept secret until deadline
- Task: predict 100 points and error bars

LASER CONTINUATION WITH LOCAL ERROR BARS



PREDICTABILITY VARIES IN STATE SPACE



2. LOCAL LYAPOUNOV "COEFFICIENT"

- **We have trained a network**
- **This can be viewed as a "skeleton"**

  It captures the deterministic part of the dynamics.

  The stochastic part is removed

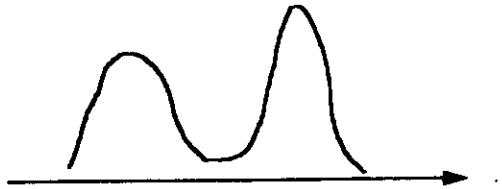
- **We can now estimate the local divergence from the derivatives**
- **This describes the local divergence (as a function of the point in state space)**
- **It contains much more information than averaging over the attractor (as typically done for the Lyapounov coefficient)**

3. PREDICTING PROBABILITY DENSITY

with Ashok Srivastava

ftp.cs.colorado.edu:/Time-Series/prob-density.ps

- **multimodal distribution**



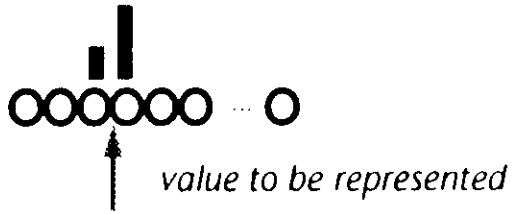
its mean is not a good description

- **representation**
  - **histogram?**

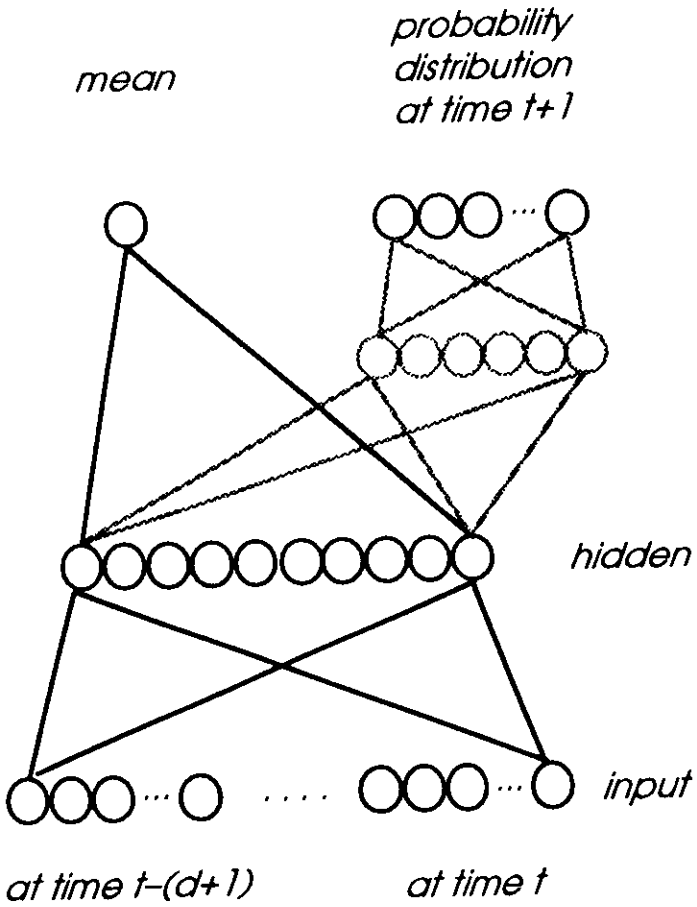
    suboptimal resolution for the predicted value (given by bin size)

  - **use fractional binning (*soft* histogram)**

    the target value (real-number) is distributed over two adjacent bins



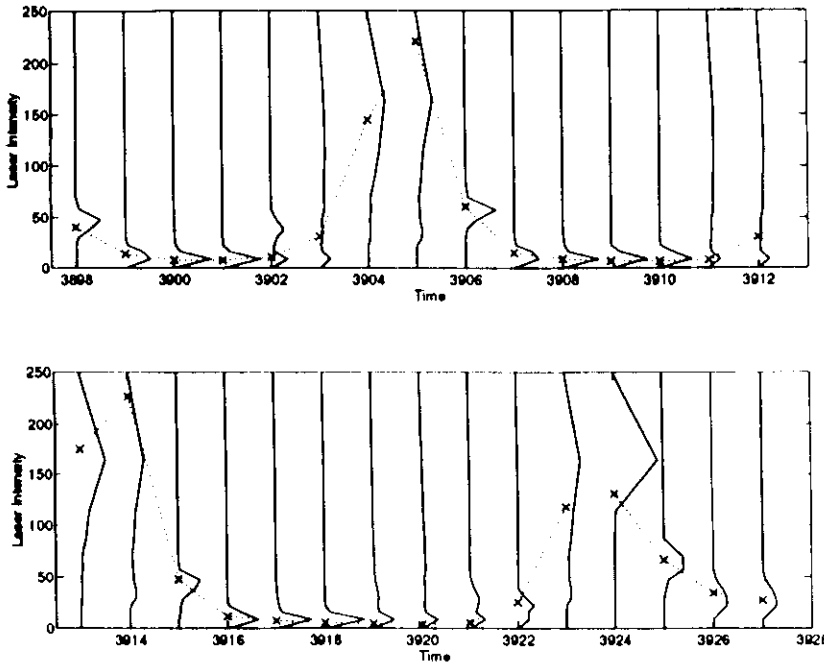
ARCHITECTURE FOR PROBABILITY DENSITY



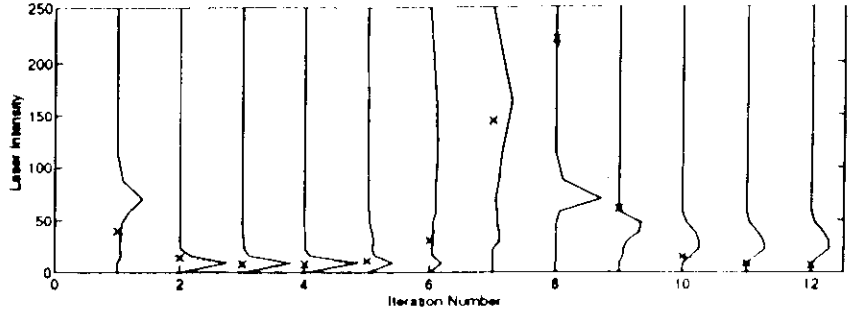
TRICKS FOR PROBABILITY DENSITY

- Some tricks to make it work
  - predict the mean in addition to density
    to find a good representation in the set of shared hidden units
  - equal mass bins
    not fixed binsize
  - stochastic teacher forcing (annealing)
    - training begins with exact values at inputs
    - gradually replace them with the predictions
- Note that iteration is straightforward
- Place in the space of processes
  - "noisefree Markov with metric"
- Example: Santa Fe laser data

SINGLE-STEP PREDICTIONS



ITERATED PREDICTIONS



INTERMEZZO: SOURCES OF UNCERTAINTY (NOISE)

stochasticity

outside shocks

chaos

divergence of nearby trajectories

sampling noise

particularlly important for small data sets

model misspecification

- **"true model" not in model space**
- **parameters not estimated well**
- **wrong error model**

GENERALIZATION ≠ MEMORIZATION

main goal: how good on future data?

Standard procedure to estimate generalization performance:

- **Split available data in three sets**
  - **training set (to estimate parameters)**
    - **fitting or approximation error**
    - **In-sample-error**
  - **cross-validation set (to estimate stopping)**
  - **test set (to estimate performance)**
    - **generalization error**
    - **out-of-sample error**

Question:

How large is the effect of this splitting? ("sampling noise")

4. ESTIMATE SAMPLING NOISE

with Blake LeBaron

ftp.cs.colorado.edu:/Time-Series/bootstrap.ps

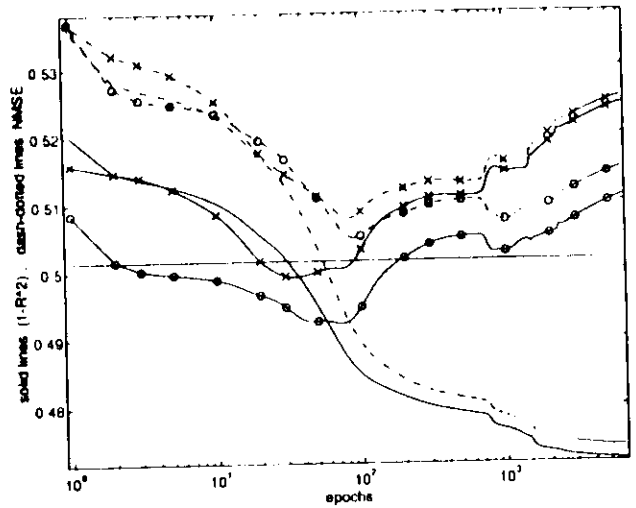
- **Bootstrap the split of the data**
  - **bootstrap test/cross-validation/test sets, then train networks on single-step prediction, and make histogram**
- **Obtain distribution of central value**
- **Results**
  - **On predictions of NYSE volume data: predict about 50% of the variance**
  - **Splitting more important for performance than initial conditions of network**
  - **No improvement over linear models**

DATA SET: DAILY VOLUME FROM NEW YORK STOCK EXCHANGE

- **Data**
  - **December 3, 1962 ... September 16, 1987**
    6230 days
  - **Total trading volume on NYSE**
  - **Dow Jones Industrial Index**
- **Network**
  - **Inputs: 3 past values each of**
    - **volume**
      log[turnover / 100-day average of turnover]
    - **daily returns, and their absolute values**
      return = log[price(today) / price(yesterday)]
    - **log(volatility)**
      volatilility: exponential decayed squared returns
  - **Hidden units**
    2 ... 10 tanh
  - **Output**
    volume
  - **Cost function: squared error**
    (prediction– actual value)^2

LEARNING CURVES

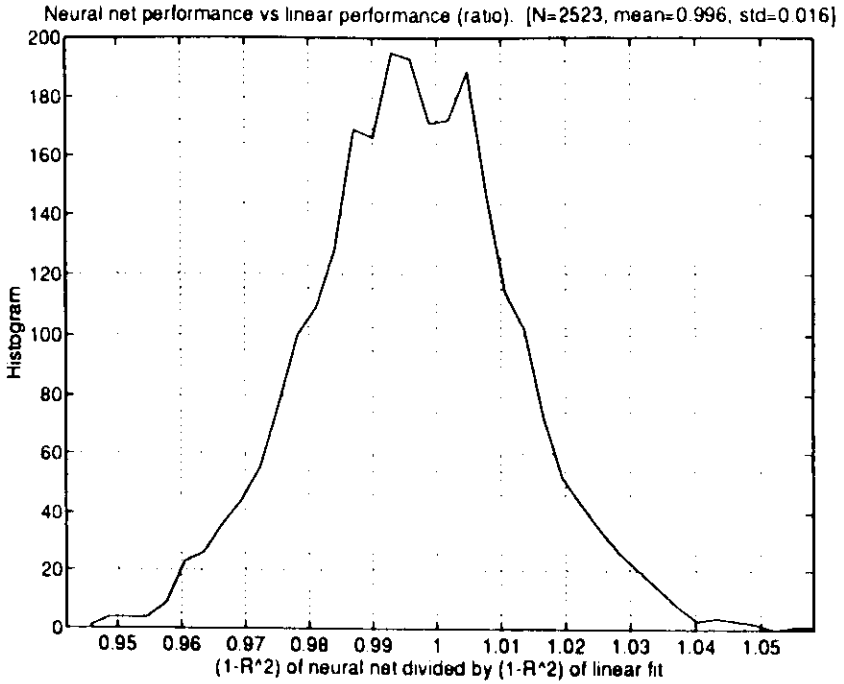
- **Plot error** (for one run) **as function of the number of epochs** (passes through the data)



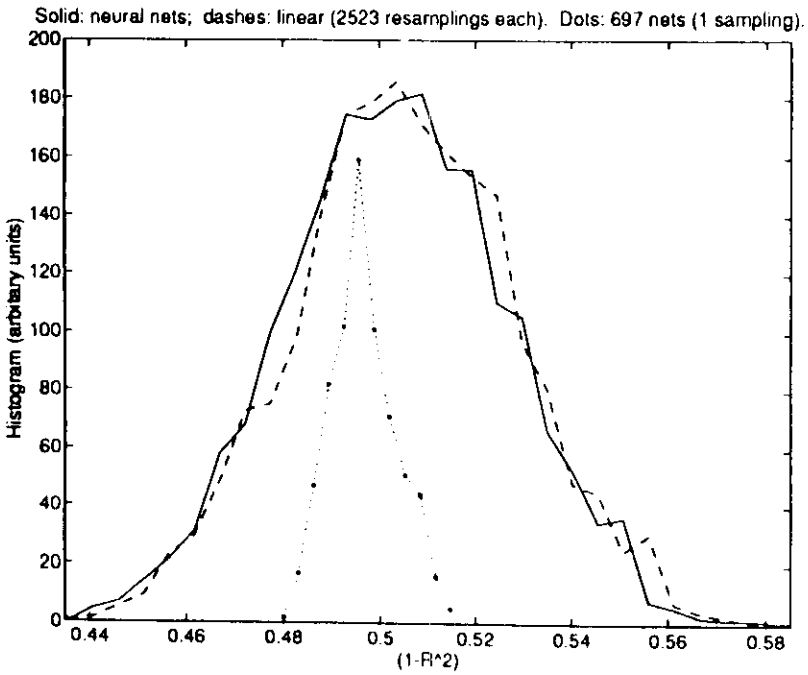
  - Training error (——) decreases monotonically
  - *Overfitting*: Cross-validation error (x) and test error (o) first dereases, then increase
  - Network extracts features from training data (in-sample) that do not generalize to new data (out-of-sample)
  - early stopping: use network at that epoch with minimum on cross-validation set

COMPARISON TO LINEAR MODEL

- **Plot ratio of network performance compared to performance of liner fit**

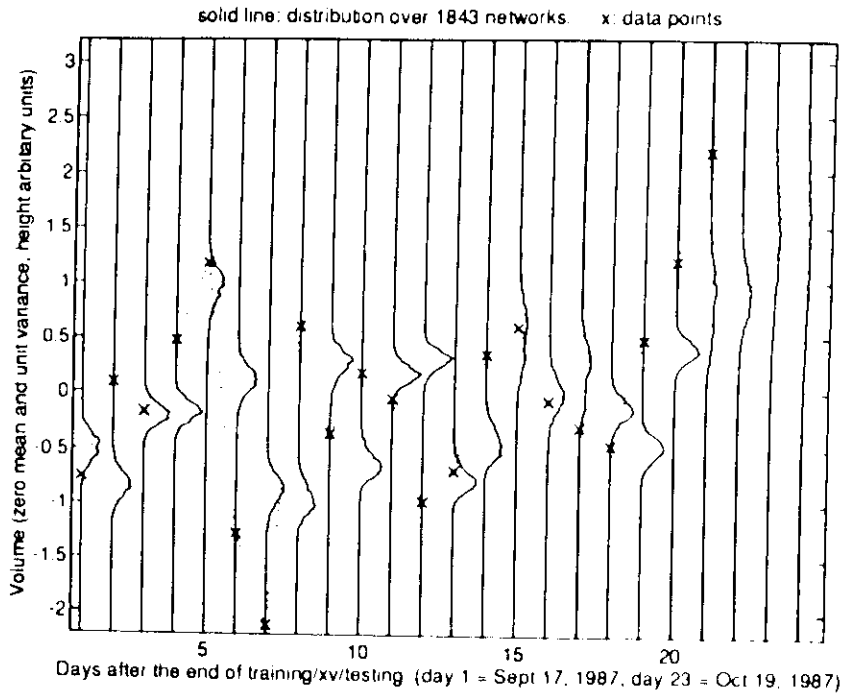


EFFECT OF RANDOM SPLITS VS NETWORK INITIALIZATION



PROBABILITY DISTRIBUTION

- Captures uncertainty from splits
- Time period includes 1987 crash



ENTICING NONLINEARITIES

Simple network no better than linear model

How to entice net to use nonlinearities?

Change task:

- **Train on residuals of linear fit**

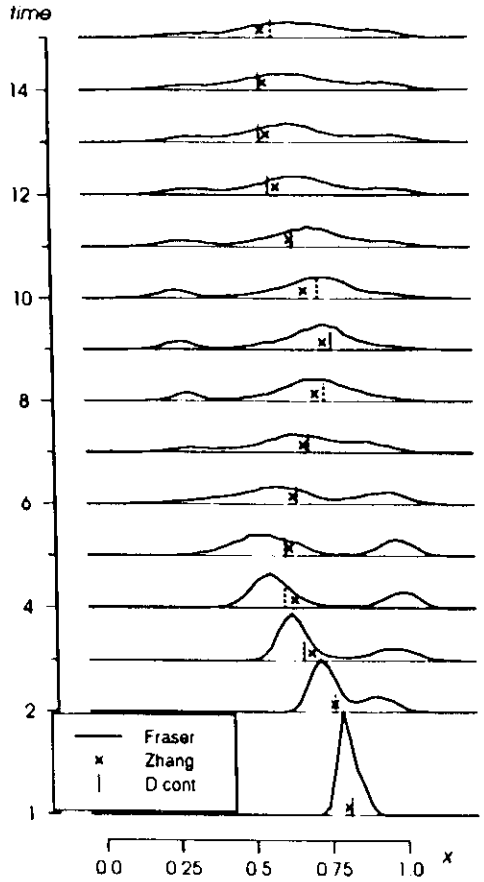
Change representation:

- **Thermometer code**
- **Place coding; Fractional binning**
- **Make it a classification task**
- **Predict quantiles**
- **Use error-correcting codes (blow up into larger space with nonlinear combinations)**

What else??

5. GENERAL MONTE CARLO

- **Add noise to inputs to obtain confidence intervals**
- **Example: Fraser and Dimitriadis (1994)**



SUMMER 1994: HU BERLIN AND CU BOULDER

both accessible through World Wide Web (WWW)

Humboldt Universität zu Berlin (Wirtschaftsinformatik)

Methodenbank

Client - server architecture

Combines techniques from several communities (e.g., economics, finance, physics, statistics, neural networks,...)

Model management

Importance of interactive exploratory analysis (visualization)

University of Colorado at Boulder (Computer Science)

Data Set Citation Index

Gatherers - broker architecture

For each benchmark time series, it maintains hyperlinks to url's of papers that use / analyze / predict the data

CU BOULDER AND HU BERLIN

Need / Chance for meta-analysis:

- Can't do much analytical (yet), need to do as much empirical as possible:
- Both projects allow the collection of "meta data": monitor user behavior to learn more about interplay between methods and data

DON'T TRY THIS A HOME... CALL A CONNECTIONIST!

or get the papers of this talk:

- via WWW / Mosaic:

http://www.cs.colorado.edu/homes/andreas/public_html/Home.html

- or via ftp:

ftp.cs.colorado.edu (128.138.243.151)

SFI-book.bibliography
error-bars.ps
prob-density.ps
bootstrap.ps
...

FURTHER INFORMATION

- Book

edited by A.S.Weigend and N.A.Gershenfeld
Addison-Wesley (1994) 672 pages, 800 references

ISBN 0-201-62602-0 (pb, $32.25)
ISBN 0-201-62601-2 (hc, $49.50)

- Overview article

anonymous ftp to `ftp.cs.colorado.edu`
`Time-Series/CU-CS-670-93`

- Data

anonymous ftp to `ftp.santafe.edu`
or to `ftp.cs.colorado.edu`